Protein Interaction Prediction using Inferred Domain Interactions and Biologically-Significant Negative Dataset

Xiao-li Li, Soon-Heng Tan, and See-Kiong Ng

Institute For Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613
{xlli,soonheng,skng}@i2r.a-star.edu.sg

Abstract. Protein domains are evolutionarily-conserved structural or functional subunits in proteins that are suggestive of the proteins' propensity to interact or form a stable complex. In this paper, we propose a novel domain-based probabilistic classification method to predict protein-protein interactions. Our method learns the interacting probabilities of domain pairs based on domain pairing information derived from both experimentally-determined interacting protein pairs and carefully-chosen non-interacting protein pairs. Unlike conventional approaches that use random pairing to generate artificial non-interacting protein pairs as negative training data, we generate biologically meaningful non-interacting protein pairs based on the proteins' biological information. Such careful generation of negative training data set is shown to result in a more accurate classifier. Our classifier predicts potential interaction between any pair of proteins based on the probabilistically inferred domain interactions. Comparative results showed that our probabilistic approach is effective and outperforms other domain-based techniques for protein interaction prediction.

1 Introduction

Cellular processes are biochemical events that are typically achieved by the interactions of proteins with one another. The elucidation of protein interactions is therefore the necessary first-step for understanding the biology of cellular processes. Many experimental methods have been developed to detect protein-protein interactions, however, none of the current experimental methods is adequate to interrogate the entire interactome [11, 15]. It is therefore useful to develop complementary computational methods for predicting new protein-protein interactions.

Several computational techniques have been proposed to predict protein-protein interactions. For example, potential protein interactions can be derived from gene context analysis such as gene neighborhood [3, 13], gene fusion [5, 9], and gene co-occurrences and phylogenetic profiles [8, 14]. Alternatively, the physiochemical properties or tertiary structure of proteins can also be used for predicting interactions[1, 10].

Recently, however, there is an increased focus on using *protein domains* to predict protein-protein interactions [4, 6, 7, 12, 16]. Protein domains are evolutionarily-conserved structural or functional subunits in proteins found across different proteins. They are often found to participate in intermolecular interactions with one another. The existence of certain domains in proteins can therefore suggest the possibility of interaction between two proteins. As such, the analysis of many protein-protein interactions can be reduced to understanding the underlying domain-domain interactions between two proteins.

Domain-based protein interaction prediction methods generally consist of two main steps: 1) inferring domain-domain interactions from known protein interactions, 2) predicting protein interactions based on the inferred domain-domain interaction information. A few domain-based interaction detection techniques have recently been proposed. Deng *et al.* described a Maximum Likelihood estimation technique to infer domain-domain interactions that was then used to predict protein interactions [4]. Wan *at al.* presented a alternative statistical scoring system as a measure of the interaction probability between domains [16]. Ng *et al.* devised an integrative approach to infer the protein domain interactions [12] from other data sources in addition to experimentally determined protein interactions. Han *et al.* designed a probabilistic framework that takes domain combinations instead of single domains as basic units of protein interactions [6, 7].

These proposed techniques can be grouped into two main paradigms in terms of the way they infer domain-domain interactions. The domain interactions that are used for predicting protein-protein interactions are learned either (1) from an interacting protein set or positive class only [4, 12, 16], or (2) from both an interacting protein set and an artificially generated non-interacting protein set as negative set; the latter being generated by randomly pairing the proteins [6, 7]. In the case of (1) where learning is conducted only from an interacting protein set, many false positive domain pairs may be derived because these domain pairs may occur in the (unavailable) negative set with high frequency. In the case of (2), the use of a putative negative data set helps alleviate this problem. However, using artificially generated non-interacting protein set as negative set is inadequate for inferring domain-domain interactions because the randomly generated negative dataset may contain interacting protein pairs. In addition, if the artificially generated negative dataset is subsequently used in evaluating the performance of classifier, it will lead to inaccurate computation of the actual sensitivity and specificity of the technique.

In this paper, we propose a novel probabilistic technique to infer domain-domain interactions using both positive and negative training datasets. Our probabilistic model was able to outperform other domain-based techniques in predicting potential protein interactions. Unlike conventional approaches that use random pairing to generate artificial non-interacting protein pairs as negative training data, we generate biologically meaningful non-interacting protein pairs based on the proteins' biological information, namely, proteins are most unlikely to interact if they are from different cellular locations and functional categories.

We showed that the performance of classifier is improved with the more confident negative dataset.

2 Methods

Our proposed approach classifies a protein pair to be either interacting or non-interacting based on inferred underlying domain-domain interactions. The approach consists of three steps as follows: 1) generate the negative set N (non-interacting protein pairs); 2) infer domain-domain interactions based on the interacting proteins pair set I and the negative set N; 3) build a classifier based on the interacting probabilities of domain pairs. Below, we present the methods for these three steps in turn.

Generate the negative set: Proteins are most unlikely to interact if they are from different cellular locations and functional categories. So our generated negative set only pairs those proteins located at different locations and with different functions. Algorithm 1 shows how to generate non-interacting protein pairs (negative set).

Algorithm 1 Generate non-interacting protein pairs

```
 1: **Input**: interacting set $I$, protein set $P$;
 2: **Output**: negative set N;
 3: BEGIN
 4: Set $N = \emptyset$;
 5: for all the protein $p_i \in P$ do
 6:   Search $p_i$'s locations (l) and functional categories (c);
 7: end for
 8: Combine all protein pairs into a set $PS$: $PS = \{(p_i, p_j) | p_i \in P, p_j \in P, i \neq j\}$;
 9: repeat
10:   for each protein pair $(p_i, p_j) \in PS$ do
11:     if $(p_i, p_j) \notin I$ then
12:       if $((p_i.l \neq p_j.l) \wedge (p_i.c \neq p_j.c))$ then
13:         $N = N \cup \{(p_i, p_j)\}$;
14:       end if
15:     end if
16:     $PS = PS - \{(p_i, p_j)\}$;
17:   end for
18: until $(PS = \emptyset)$
19: END
```

In Algorithm 1, for each protein in P, the set of proteins of interest, we retrieve the biological information about its locations and functional categories (Steps 5-6) from the MIPS database[1]. Then, from Step 9 to Step 18, we check

[1] http://mips.gsf.de/genre/proj/yeast/index.jsp

each protein pair (p_i, p_j) in protein pair set PS: if it is already in the interacting protein set I , we eliminate it from PS; otherwise, if p_i and p_j are located at different cellular locations and from different functional categories, we add them into negative set N.

Note that in Step 12, a protein (p_i or p_j) may be located at multiple locations and has multiple functions. We consider (p_i, p_j) to be non-interacting only if none of p_i's locations and functions match the p_j's locations and functions. In addition, because the proteins' functional classifications given in MIPS are hierarchical, we will only regard two proteins to have different functions at the highest possible level of the MIPS functional hierarchy(Level 1). Such strict selection strategy helps us get a much purer negative set for training our classifier.

Infer domain-domain interactions: The objective of this next step is to assign interaction probabilities to each domain pair based on its occurrence in the protein-protein interacting set I and the negative set N. For a protein pair $(p_i, p_j) \in I$, we infer that domain $d_{i,r}$ potentially interacts with domain $d_{j,s}$ with a probability of $1/(|p_i| * |p_j|)$, where $|p_i|$ and $|p_j|$ are the number of domains in proteins p_i and p_j respectively; $d_{i,r}$ and $d_{j,s}$ are the r-th and s-th domains of proteins p_i and p_j respectively.

Given that a domain pair (d_x, d_y) may occur in many interacting protein pairs of I, the interacting frequency of (d_x, d_y) in I is defined as:

$$N((d_x, d_y), I) = \sum_{i=1}^{|I|} \lambda_i(d_x, d_y) * \frac{1}{|p_x^i| * |p_y^i|} \tag{1}$$

where (p_x^i, p_y^i) is the i-th protein pair in I and $\lambda_i(d_x, d_y)$ is the total number of occurrences of the domain pair (d_x, d_y) in (p_x^i, p_y^i). We compute $N((d_x, d_y), N)$, the interacting frequency of (d_x, d_y) in N, in a similar way:

$$N((d_x, d_y), N) = \sum_{i=1}^{|N|} \lambda_i(d_x, d_y) * \frac{1}{|p_x^i| * |p_y^i|} \tag{2}$$

Let a set of pre-defined classes be $C = \{I,\ N\}$ and all the domain pairs set be DP. For any domain pair $(d_x, d_y) \in DP$, their interacting probability $P((d_x, d_y)|c_e)$, with Laplacian smoothing and $c_e \in C$, is defined as:

$$P((d_x, d_y)|c_e) = \frac{1 + N((d_x, d_y), c_e)}{|DP| + \sum_{k=1}^{|C|} N((d_x, d_y), c_e)} \tag{3}$$

For a domain pair (d_x, d_y), the greater the interacting probability $P((d_x, d_y)|I)$, the more frequent it occurs in the interacting set I. However, since such a domain pair may also be chanced occurrences in class I, it is necessary to check its interacting probability in N: $P((d_x, d_y)|N)$. Obviously, if $P((d_x, d_y)|I)$ is significantly larger than $P((d_x, d_y)|N)$, then the domain pair (d_x, d_y) is likely to be a genuine domain-domain interaction. Otherwise, if $P((d_x, d_y)|N)$ is similar or even bigger than $P((d_x, d_y)|I)$, then the domain pair is unlikely to be interacting. In other words, to check if a domain pair (d_x, d_y) interacts, we compute its interacting probabilities in both interacting set I and negative set N.

Note that the purity of N can affect the accuracy of inferred domain-domain interactions. If N were generated from randomly paired proteins, the false negative protein pairs in N will result in the inference of many domain pairs that should have occurred only in interacting protein set (i.e. positive class). This will result in assigning inaccurate interacting probabilities to domain pairs and subsequently affect the accuracy of the eventual classifier to infer protein interactions.

Build a protein interaction classifier: Given a protein pair (p_i, p_j), in order to perform classification (i.e. to judge whether the proteins may interact with each other or not), we compute the posterior probability $P(c_e|(p_i, p_j)), c_e \in C$. The prior probability $P(c_e)$ of class c_e is defined as:

$$P(c_e) = \frac{\sum p(c_e, (p_i, p_j)), (p_i, p_j) \in I \cup U}{|I| + |N|} \tag{4}$$

Based on Equations (4) and (3), our proposed technique uses the joint probabilities of domain pairs and classes to estimate the probabilities of classes given a protein pair. Our classifier is described as follows:

$$P(c_e|(p_i, p_j)) = \frac{p(c_e) * \prod_{m=1}^{|p_i|*|p_j|} p((d_{i,r}, d_{j,s})|c_e)}{\sum_{k=1}^{|C|} p(c_e) * \prod_{m=1}^{|p_i|*|p_j|} p((d_{i,r}, d_{j,s})|c_e)} \tag{5}$$

For a protein pair (p_i, p_j), the class with highest $P(c_e|(p_i, p_j))$ is assigned as its final class label. In other words, if $I = argmax_{c_e} P(c_e|(p_i, p_j))$, then the protein pair (p_i, p_j) will be classified as an interacting pair. Otherwise, it is classified as non-interacting.

3 Evaluation

In this section, we evaluate the proposed technique for predicting protein interactions. Positive and negative datasets are employed to train a classifier and to evaluate the performance of our method. For positive datasets, interacting proteins are retrieved from DIP[2]—a comprehensive curated catalog of about 44,482 experimentally determined protein-protein interactions in over 110 organisms. We select all *yeast* interactions in DIP to construct our positive dataset I as this species is particularly well-studied. The *yeast* positive dataset consists of 15,658 interactions among 4,749 *yeast* proteins. The negative set of non-interacting protein pairs used in this work is constructed using Algorithm 1 described in the previous section. Proteins are paired up only if they are not from the same cellular location and functional category. This results in a very large negative set of 213,560 protein pairs. To avoid size bias between the positive and negative datasets, we randomly assembled a negative set N with the same number of protein pairs as I.

[2] http://dip.doe-mbi.ucla.edu/dip/

The domain information of proteins are obtained from the **Pfam** database [2], which contains a large collection of multiple sequence alignments and profile hidden Markov models of protein domains. Both **Pfam-A** and **Pfam-B** are used to ensure sufficient coverage. We first infer the domain-domain interactions from both positive set I and negative set N. Each domain pair gets an interacting probability for I and N using Equation (3).

Table 1. Top 10 interacting and non-interacting domain pairs

Interacting domain pairs	Non-interacting domain pairs
(PF07719, PF00515)	(PF00153, PF00400)
(PF02985, PF02985)	(PF00560, PF00172)
(PF00515, PF00515)	(PF00137, PF00400)
(PF00400, PF00118)	(PF00172, PF07714)
(PF07719, PF07719)	(PF00023, PF00153)
(PF02985, PF00514)	(PF00036, PF00400)
(PF00400, PF00514)	(PF00560, PF00096)
(PF00036, PF00612)	(PF00400, PF00122)
(PF00076, PF00514)	(PF00702, PF00400)
(PF00432, PF01239)	(PF04082, PF00153)

For illustration, Table 1 shows the top 10 interacting and top 10 non-interacting domain pairs respectively. The top interacting domain pairs have maximal values of $P((d_x, d_y)|I)/P((d_x, d_y)|N)$. In other words, these are domain pairs with biggest $P((d_x, d_y)|I)$, while smallest $P((d_x, d_y)|N)$. The top non-interacting domain pairs show those with significant occurrence in non-interacting protein pairs. As we know, not all domain pairs derived from protein-protein interactions are truly interacting as some could occur in interacting proteins by chance. This could lead to false positive domain-domain predictions if we learn from the positive class I only. For example, domain pair (PF07714, PF00400) and (PF00515, PF00806) occurred 170 and 70 times in I respectively. If we just learn from I, it is natural to infer them to be interacting domain pairs since they have high occurrence in interacting set. However, with the help of our biological refined negative class N, we were able to eliminate them since both domain pairs also occurred 1141 and 160 times in N respectively. Furthermore, since our negative class N is more biologically significant than randomly paired proteins, we can estimate the interacting probabilities of each domain pair more precisely and thus result in a more accurate classifier.

For evaluation, we use the inferred domain-domain interactions to classify protein pairs. A 5-fold cross validation is performed to test the accuracy of the classifier described in Equation (5). We compare our results with the reported results using the "Hybrid Classification" technique from reference [7] and the "Possibility Ranking" technique from reference [6], both of which used positive and negative training datasets for improved protein interaction prediction. Table 2 shows the comparison results of four domain-based protein predication techniques in terms of sensitivity and specificity. The first two techniques were from references [7] and [6]. The other two are our probabilistic technique with two different negative sets, namely, randomly paired negative set (random pairs) and the biologically significant negative set (biological refinement).

Table 2. Classification results of different techniques

Techniques	Specificity	Sensitivity
Hybrid Classification [7]	56.00	86.00
Possibility Ranking [6]	75.00	84.36
Our technique with random pairs	83.71	84.80
Our technique with biological refinement	90.21	87.52

Compared with the techniques in [7] and [6], our probabilistic technique was able to achieve much higher specificity at similar sensitivity regardless of whether it has been trained with random protein pairs as the negative set or the refined negative set assembled using biological domain knowledge. Our classifier that was trained with the biologically refined negative dataset gave the best performance, obtaining an increase of 6.5% and 3.3% in specificity and sensitivity respectively as compared to the same probabilistic classifier trained with negative dataset of randomly paired proteins. This shows that the use of biological domain knowledge for negative dataset construction can benefit the prediction performance of the eventual classifier built on the training data.

The techniques from [7] and [6] were not tested with cross-validation. They randomly selected 20% DIP data as test set and the remaining 80% as training set. Then they repeated their experiments 3 times and got the average results. In fact, as reported in [7], their specificities was rather fluctuating according to the selected test sets. Our method is more robust as our results fluctuated only within 3% in each division of cross validation.

Table 3. Performance of classifier with the different size of N

Size ratio	2	4	6	8	10
Specificity	93.00	95.51	96.07	96.27	96.32
Sensitivity	83.21	75.54	73.14	71.50	71.30

Finally, we also investigate how the size of negative set may affect the performance of our classifier. We systematically increase the size of N by 2 to 10 times. The results are shown in Table 3. With the increase in the size of the negative set, the specificity of our classifier increases while the sensitivity decreases.

One reason that sensitivity has decreased is that the imbalance of positive and negative training set makes our classifier biased towards the negative class N. However, we believe that it is possible to get better performance through intelligently selecting negatives, and we will leave this problem as our future study. Another possible future work involves integrate other biological features such as "amino acid composition" with the domain information in the prediction of protein interactions.

4 Conclusion

In this paper, we predict protein-protein interactions based on domain information. Our learning algorithm first constructs a biologically meaningful negative set based on biological domain knowledge. It then infers the underlying domain interactions based on their probabilities in both interacting class and non-interacting class. A probabilistic classifier for predicting protein interactions is then built upon the inferred probabilistic domain interactions. Our experimental results show that our probabilistic approach is effective and outperforms other similar domain-based techniques for protein interaction prediction.

References

1. J. R. Bock and D.A. Gough. Prediction of protein-protein interaction from primary structure. *Bioinformatics*, 17:455–460, 2001.
2. F. Corpet, F. Servant, J. Gouzy, and D. Kahn. Prodom and prodom-cg: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res*, 28:267–269, 2000.
3. T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, 23:324–328, 1998.
4. M. Deng, F. Sun S. Metha, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome research*, 12:1540–1548, 2002.
5. A. J. Enright, I. Illiopoulos, N. C. Kyrpides, and C.A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90, 1999.
6. D. Han, H. Kim, W. Jang, and S. Lee. Domain combination based protein-protein interaction possibility ranking method. In *IEEE Fourth Symposium on Bioinformatics and Bioengineering (BIBE2004)*, pages 434–441, 2004.
7. D. Han, H. Kim, J. Seo, and W. Jang. Domain combination based probabilistic framework for protein-protein interaction predication. *Genome Informatics*, 14:250–259, 2003.
8. M. A. Huynen and P. Bock. Measuring genome evolution. *Natl Acad. Sci*, 95:5849–5856, 1998.
9. E.M. Marcotte. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285:751–753, 1999.
10. S. Martin, D. Roe, and J.L. Faulon. Predicting protein-protein interactions using signature products. *Bioinformatics*, 20:1–9, 2004.
11. S.K. Ng and S.H. Tan. Discovering protein-protein interactions. *Journal of Bioinformatics and Computational Biology*, 1(4):711–741, 2004.

12. S.K. Ng, Z. Zhang, and S.H. Tan. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19:923–929, 2003.
13. R. Overbeek, M. Fonstein, M. D'Souza, G.D. Pusch, and N.Maltsev. The use of gene clusters to infer functional coupling. *Natl Acad. Sci.*, 96:2896–2901, 1999.
14. M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeastes. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Natl Acad. Sci*, 96:4285–4288, 1999.
15. C. von Mering, R. Krause, B. Snel, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.
16. K. K. Wan and P. Jong. Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair. *Genome Informatics*, 13:45–50, 2002.